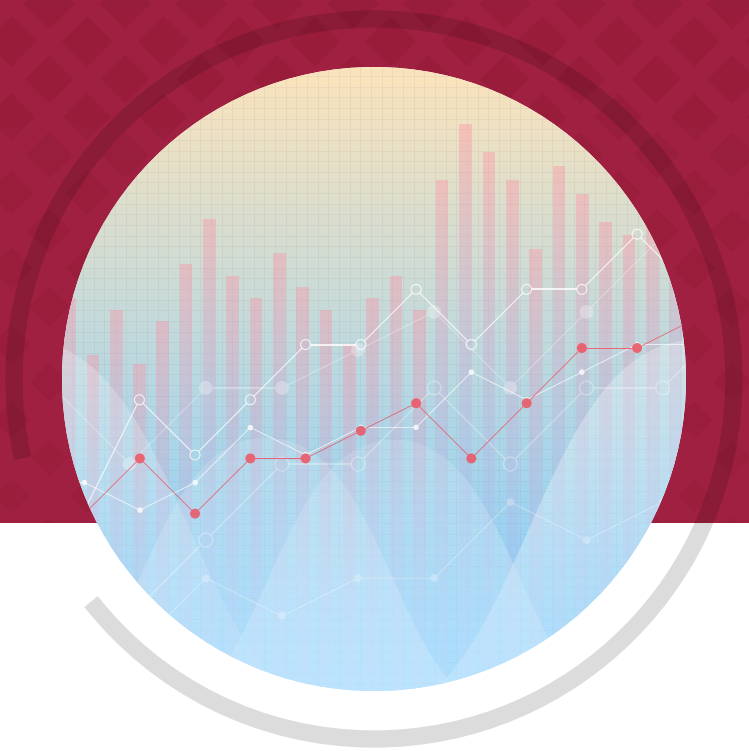
On Growth Models, Time for California to Show Some Improvement

Morgan Polikoff



California is one of just two states (with Kansas) that does not use a student-level growth model to measure school performance. This brief lays out a number of common beliefs about growth models and provides evidence that these beliefs are inaccurate or unsupported. In so doing, the brief makes a positive case that the state should adopt such a model and replace the current "change" metric in the California School Dashboard. Two specific models—student-growth percentiles and residual-gain growth models—would be a dramatic improvement over what the state currently uses and would much more validly identify schools succeeding and in need of support.

September 2019

PACE
Policy Analysis for California Education

Introduction

Educational accountability in California is in a new era. As the federal government has relaxed its requirements for consequential accountability, the state has taken a new approach to evaluating and supporting schools. In the past six years, California:

1. Retired the Academic Performance Index, which it had been using to rate schools for nearly 15 years.
2. Ended No Child Left Behind accountability, moving from a focus on rewards and sanctions to a model of continuous improvement.
3. Enacted the Local Control Funding Formula and Local Control Accountability Plans.
4. Rolled out the new California School Dashboard to report on school performance on multiple test-based and non-test indicators.

A look at the key indicators in the Dashboard illustrates California's effort to consider not just the *status* of school performance, but also *changes* in school performance. Specifically, California has chosen to include in the Dashboard ratings a "cohort-change" model that compares this year's average score in a school or district to last year's average score.

For many reasons (which I describe below), it is important for the state to include a measure of growth in its accountability system. But is the cohort-change model the right choice? California is one of just two states (the other is Kansas) that does not calculate or report a *student-level growth model* (i.e., one based on comparing the growth in achievement of individual students from year to year). Should it? What are some of the key considerations in selecting a growth model, and how might California use what we already know from other states and from decades of research to make the best selection?

The purpose of this brief is to discuss the reasons why California should adopt a measure of student growth that aligns with what we know about the design of such measures and their use in accountability and continuous improvement systems. To do this, the brief presents a number of common misconceptions about growth models and dispels them using existing evidence. For the most part, when this brief talks about student growth models, it is referring to the Residual Gain Model[1]; when other models are discussed they are called out as such.

Growth Model Misconceptions

California is too different from other states to learn about growth models from research done elsewhere

Many people believe that technical research, such as growth-model research, must be context-specific to be relevant. The truth is that there is no reason to believe that the general findings from the technical literature on growth models are context-specific. Furthermore, growth-model research comes from a wide variety of contexts, many of which look like California.

There is a very large body of research on growth models[2], and more research is being produced all the time. This research comes from states that look demographically similar to California (e.g., Texas[3] and Florida[4]) and from states that look quite different (e.g., Tennessee[5] and Missouri[6]). None of the recent reviews on the topic, nor any of the individual state-specific studies, provide any indication that the methodological recommendations they make are state- or context-specific.

Furthermore, though California often thinks of itself as being distinct from other states, there are many ways in which California is similar to other large, diverse states. Although California is the most populous state (enrolling approximately 6.2 million students), it is only 17 percent larger than Texas in student enrollment[7]. California is highly diverse, with only 24 percent white students and 59.8 percent African American and Latinx students, but Texas is just 28.5 percent white and 64.8 percent African American and Latinx. California has a large number of school districts (1,059), but Illinois has 970 and Texas has 1,241. California also has a large percentage of students eligible for free- and reduced-price lunch (58.9 percent), but Florida has 58.8 percent. In short, California looks very much like other large states along most any dimension. There is no reason to think California cannot learn about growth models, their designs, and their effects by drawing on data and lessons from other states.

California already has a growth model in the Dashboard

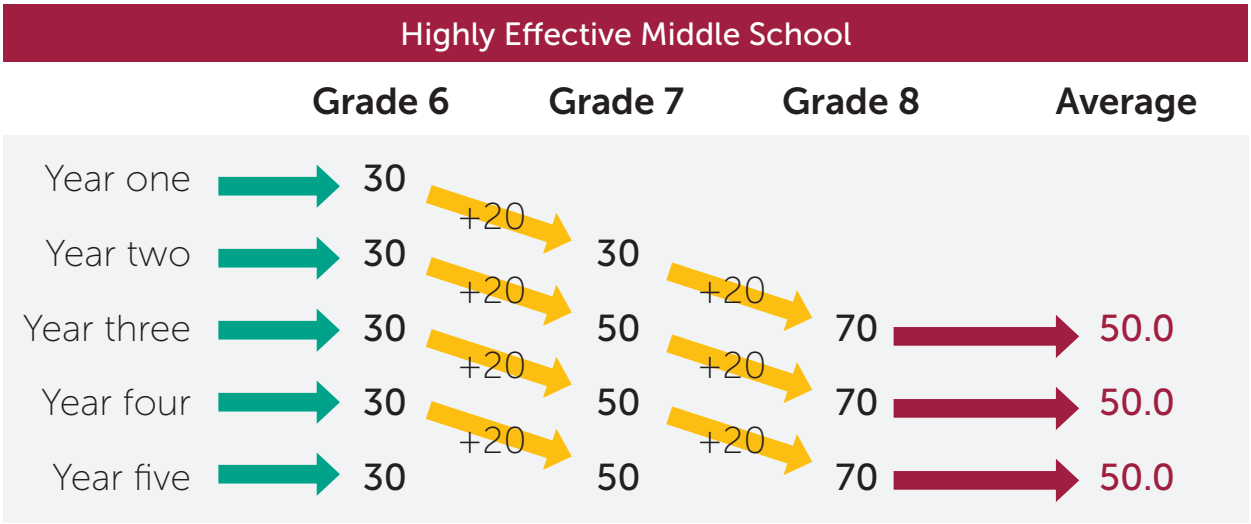
Many people might look at the California School Dashboard and see that it accounts for a school's "change" from last year to this year in assigning a Dashboard rating; they might believe that this change score is equivalent to a growth model. The truth is that the change score is not a growth model—it is more commonly referred to as a "cohort change" model or an "improvement" model[8]. It is substantively different from (and inferior to) a growth model.

The difference between a growth model and this cohort change model is straightforward. A growth model tracks the performance of individual children from year to year, comparing the growth rates of children in different classrooms, schools, or districts.

In contrast, a cohort-change model like the one California currently uses compares this year's students in a school or district to last year's students. They are fundamentally different approaches to looking at the change in performance over time (they literally measure different things), so the current Dashboard measure is not a growth model.

Consider a middle school where the students come in at the 30th percentile in 6th grade, advance to the 50th percentile in 7th grade and the 70th percentile in 8th grade[9]. This school is doing *phenomenal* things for children—raising their achievement dramatically. A growth model, shown in Table 1, would reflect this impact and would show this school as a huge positive outlier. In contrast, if the arriving 6th graders stayed at the 30th percentile year after year (as they likely would, given the stable relationship of school-average poverty with achievement levels), the state's cohort change model would show this school as middle-of-the-road.

Figure 1. How cohort change models fail to measure student growth



Source: Albert Shanker Institute

Even if the current change measure isn't the same as a student-growth model, it tells us the same things

Many people might believe that the state's cohort-change model and a true growth model might differ technically while still producing the same or very similar results. The truth is that cohort-change models and adjusted-gain growth models can (and often do) produce substantially different results because they measure fundamentally different things

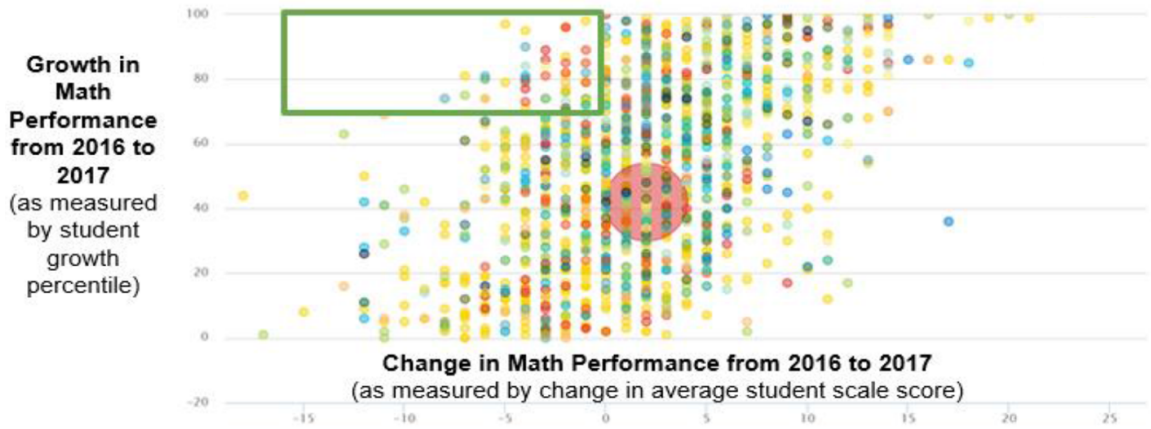
There are many reasons why these models produce different results. One obvious reason is because cohorts of students can vary dramatically from year to year, so the groups of students being compared to one another in a cohort-change model are often

very different from one another. While comprehensive California data are not available, it's estimated that around 8-15 percent of children move out of a school or district in a given year (numbers are even greater in high-needs schools)[10].

At a more basic level, it is clear to see that a school that causes student achievement to grow substantially from year to year (thus, performing well under a growth model), could have no or little change in the cohort-change model if the school enrolled similar kinds of students from year to year. The opposite is also true—a school in a gentrifying area could enroll more affluent children each year, causing it to look better in a cohort-change model even if it is not actually causing student achievement to grow.

To see how pronounced the differences in the approaches are, a recent analysis using data from the CORE districts compared school ratings based on the cohort-change model currently used to growth model results based on a student-level growth model[11]. The results, seen in Figure 2, showed that large proportions of schools identified as low-performing using a cohort-change model were actually high-growth schools (the top left corner). Similarly, many schools that were above average on the cohort-change model scored below average on the growth model (the bottom right corner). In this instance, it is not that both models are equally wrong, it's that the cohort change model is giving more incorrect (i.e., less valid) signals about school performance.

Figure 2. Comparison of cohort-change model (X axis) and student growth percentile (y axis)



Source: CORE Districts. 2018 analysis of residual gain student-level growth models.

There's no agreement among researchers on which growth models to use

Many people might think there are too many kinds of growth models with no agreement even among experts about which of the models is the best. The truth is that only a very small number of models are regarded as the "gold standard," and the choice of one model over another is more about values and intended uses than it is about which model is the best.

Under the Every Student Succeeds Act, 42 states are using just four kinds of models: a student-growth percentile model, a value table, a growth-to-standard approach, or a residual- gain/value-added model[12]. Of these, 31 states are using one or both of the closely-related[13] student growth percentile and residual gain/value-added models. A variety of studies support the general conclusion that these kinds of models—regression-based models that determine how much better or worse children score on a test given their prior achievement (and possibly other variables)—are the most appropriate for making inferences about schools' effects on student achievement[14]. Residual-gain models fare the strongest from a validity standpoint, while student- growth percentile models fare slightly worse on validity but may be more understandable by parents or educators. There are numerous available reviews of the evidence about the strengths and weaknesses of different models, but there is broad consensus among researchers who study growth models that these two approaches are the least biased and most accurate.

Controlling for student demographics in a growth model means we are setting different goals for different children

Some residual-gain models incorporate demographic information about students, including possibly their free- and reduced-price lunch status, EL status, disability status, etc. Many people might see models that include these predictors and interpret them to mean that the model is setting different targets for students based on these demographic variables. The truth is that models of this sort *do* compare children to other children with similar characteristics and prior achievement, but the decision about whether or not to control for these characteristics in the model is a discussion about values, not a technical consideration.

There are several key questions the state might consider in deciding whether to control for student demographics in their residual-gain models. One question is whether the state is interested in comparing schools that are similar in terms of the kinds of students they serve. Put another way, should schools be punished or rewarded based on who the children are who happen to enroll at the school, or should comparisons be based on schools' effects on those students' performance? Similarly, should schools be compared fairly with themselves over time? For example, schools in rapidly gentrifying urban areas might quickly appear to be more effective because their student body is becoming more affluent. Controlling for student demographics would ensure schools are not benefiting from or being punished for demographic changes that are out of their control.

If the state is indeed interested in comparing schools net of their students' characteristics, then a second question is whether the state is more comfortable under-correcting or under-correcting for student characteristics. The answer to this question could inform the specific type of residual-gain model chosen[15]. These choices are mostly

conceptual and value-driven, however, the actual differences in the performance of these models are modest and depend on the particular covariates used and how they are included in the model[16].

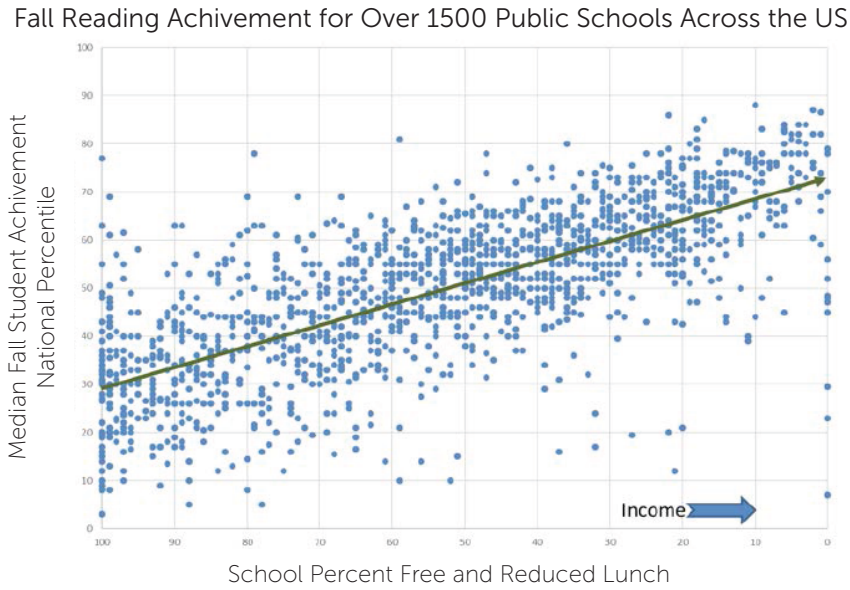
"VAM is a sham": These models don't provide school effectiveness data that could help us make valid judgments about school effectiveness

Many people have heard that residual-gain and other growth-based approaches to measuring effectiveness are biased and invalid. The truth is that the best and most recent research concludes that growth-based estimates of effectiveness using residual-gain models demonstrate *real* and *educationally meaningful* differences in effectiveness that persist for many years. Furthermore, the most sophisticated residual gain models can all but eliminate concerns about bias. Finally, the question of validity and bias is a question about the use of the results—if the results are to be used for continuous improvement and other low-stakes purposes, validity and bias concerns are dramatically reduced anyway.

Two kinds of recent studies provide evidence that the best residual gain models produce either unbiased estimates of impacts on student achievement or estimates with extremely small bias. While these studies are in the context of teacher value-added, not school value-added (there has been much less research on the latter), there is no reason to think that the general findings do not apply. In one kind of study, students are randomly assigned to teachers in order to estimate teachers' true impact on student achievement and compare it to estimates calculated from longitudinal data obtained prior to the random assignment[17]. In another kind of study, researchers use large-scale longitudinal data to relate estimates of teacher effects to students' long-term outcomes[18]. The conclusion from these studies is that "estimates of teacher value-added from standard models are not meaningfully biased by student-teacher sorting along observed or unobserved dimensions[19]."

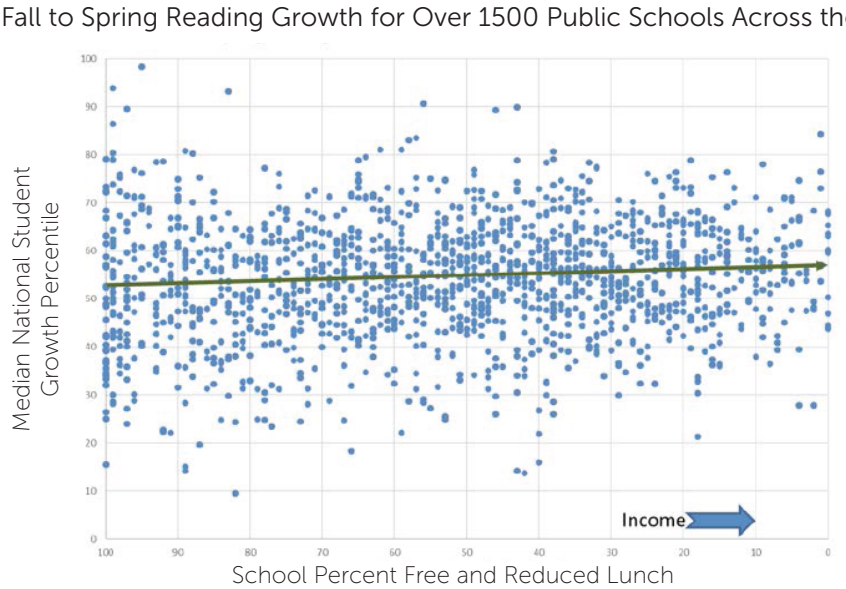
Another important dimension of the validity question is "compared to what?" Currently, schools are evaluated based on their performance levels and the aforementioned cohort-change score. A wide variety of evidence makes clear that performance levels are largely a measure of who enrolls in a school (poverty and other demographic characteristics) and have little to nothing to do with school effectiveness. Figure 3[20] demonstrates this, showing the very strong relationship between school percent free and reduced lunch and school-average achievement levels (Figure 4[21] shows that this relationship is almost nonexistent for student-growth percentiles). And, as discussed above, cohort change measures have highly questionable validity as measures of effectiveness. Thus, even if there are modest questions about bias and validity with respect to the use of residual-gain growth models to identify the effectiveness of schools, there is absolute certainty these models are better from a bias and validity perspective than what the state currently uses.

Figure 3. Student achievement levels and school average income in a large national sample



Source: Evaluating the relationships between poverty and school performance

Figure 4. Student achievement levels and school average income in a large national sample



Source: Evaluating the relationships between poverty and school performance

Growth models are too technical for educators or parents to understand

Many people might believe that growth models—especially residual-gain models that require statistical modeling—are too technical for educators or parents to understand or make use of. The truth is that educators and parents value growth data, and we have learned a great deal about how to present these data to stakeholders in ways that they understand and can use.

The first important fact is that educational stakeholders value growth data. In fact, recent research examining how parents make judgments about school quality found that parents place more weight on student growth than they do on achievement levels or any other criterion when comparing schools to make judgments about school quality[22]. Teachers also often value these data and, when properly supported and trained in how to analyze them, can use them to improve teaching and learning[23].

Not only do these stakeholder groups value growth data, but they can be supported to help understand and make use of these data. There are numerous practitioner- and parent-oriented materials available to help users make sense of these data, ranging from whole books[24] to reports[25] to short briefs[26]. It is true that growth data, especially those based on residual gain and other advanced statistical approaches, are complicated. But experience makes clear that everyone can be made to understand these data.

Growth models don't make sense in an accountability and continuous improvement system

Many people think that the data we already have in the Dashboard is sufficient for California's continuous improvement efforts. The truth is that the existing data, especially the cohort-change data, are insufficient for the task of contributing to continuous improvement.

In order for continuous improvement to succeed, we must first have an accurate sense of how schools are doing, where areas of need are located, and what practices predict improvements in outcomes. Simply put, we cannot have an accurate sense of any of these things if we do not have accurate growth data, and the current Dashboard measures do not provide accurate data on schools' effects on student learning.

Using a growth model puts teachers at risk of being fired

Many people think that because growth models have been part of high-stakes teacher evaluation systems in other states, their use in California's school accountability would lead to teacher evaluation reforms here. The truth is that creating a growth model has no bearing at all on the policy decision of how growth-model data are to be used. Furthermore, there is absolutely no indication that high-stakes teacher evaluation is on the

policy agenda in Sacramento, especially given California was one of the very few states that successfully resisted Obama-era encouragement to establish these systems.

Recommendations

Based on the existing literature and an examination of California's own goals for the Dashboard and the continuous improvement system, the state should adopt a student-level growth model as soon as possible. Forty-eight states have already done so; there is no reason for California to hang back with Kansas while other states use growth data to improve their schools.

As described above, there are just a few kinds of growth models that are used in most states. Of these, there are two possible models that are most suited to California: student-growth percentiles (SGPs) and residual-gain models. Residual-gain models can be further categorized as one-step or two-step models, as well as models that do and don't control for demographics other than prior achievement. A full review of these models is available elsewhere[27], but this brief concludes with a short discussion of the strengths and weaknesses of these approaches.

Student Growth Percentiles

Student-growth percentiles are the most widely used growth models in state accountability systems. Student-growth percentiles use students' prior test score history to answer the question "How well is this student doing this year compared to students with similar prior test scores?" SGPs are typically expressed in percentiles, so a score of 80 means that a student is doing better than 80% of students with her similar prior test history. SGPs can be averaged and reported for schools or districts. A major advantage of this approach relative to residual-gain models seems to be that it is relatively easier for practitioners to understand insofar as the numbers have a clear meaning. Disadvantages seem to be that the model may be subject to a bit more bias than residual gain models and that it has not been studied as much, so its properties are generally less well known. This model also has some technical downsides relating to its assumptions, but these are fairly typical of all approaches to measuring growth.

Residual Gain Models

SGPs are actually special cases of residual-gain models, which are also widely used by states. These models use students' prior achievement, sometimes with additional demographic or other covariates, to answer the question "How far above or below expectation is this student performing given her prior achievement (and perhaps also given her demographics and that of the school)?" Residual gain models are by far the most common models in research—they are often used, for instance, in experimental evaluations of the impact of a given treatment when student achievement is also

measured. The research consensus is that these models (especially certain kinds of residual gain models) exhibit the least bias of all available growth models. They have similar technical limitations to student-growth percentiles, and they also may be somewhat more difficult to explain because they do not produce results on the percentile scale (though they can in fact be reported on a similar percentile scale to SGPs).

A separate decision, if a residual gain model is chosen, is whether or not to adjust the model for student demographics. Most researchers who study value-added models, and are concerned most about bias and the incentives inherent in choosing a model, would prefer a model that does control for students' individual and peer demographics in addition to prior test scores. For example, Castellano and Ho (2013) argue, "If it seems that more grades [of prior achievement data] allow for an improved definition of academic peers, then why not improve the definition further by including demographic variables?[28]" Koedel and colleagues (2015) similarly argue that "in policy applications it may be desirable to include demographic and socioeconomic controls in [residual-gain models], despite their limited impact on the whole, in order to guard against [schools] in the most disparate circumstances being systematically identified as over- or under-performing.[29]" However, Koedel and colleagues also argue that the practical significance of not controlling for demographic variables is likely small—the correlation between results from models that do and do not control for demographic variables is typically close to 1.

Conclusion

Many people think that California's current Dashboard and continuous improvement system represent a dramatic change over what it replaced. The truth is that California could very easily, and at close to zero cost, choose a growth model that would represent a dramatic improvement over what currently exists. While the residual-gain model is the consensus choice of most experts, even the student-growth percentile would be a fine choice. Either way, there is more than enough information for leaders in the state, including the State Board of Education, to make a decision, and they should act now.

Author Biography

Morgan Polikoff, Ph.D., is an associate professor of education policy at the USC Rossier School of Education, the co-director of the Center on Education Policy, Equity and Governance; and the co-editor of *Educational Evaluation and Policy Analysis.* He studies the design, implementation, and effects of standards, assessment, and accountability policies.

Endnotes

1 See Castellano, K. E., & Ho, A.D. (2013). *A practitioner's guide to growth models.* Washington, DC: Council of ChiefState School Officers. Retrieved from https://scholar.harvard.edu/files/andrewho/files/a_pracitioners_guide_to_growth_models.pdf

2 For a nice overview see Castellano, K. E., & Ho, A.D. (2013). Ibid. Also see Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review,* 47, 180-195. doi.org/10.1016/j.econedurev.2015.01.006

3 Lincove, J. A., Osborne, C., Dillon, A., & Mills, N. (2014). The politics and statistics of value-added modeling for accountability of teacher preparation programs. *Journal of Teacher Education,* 65(1), 24-38. doi.org/10.1177/0022487113504108

4 Winters, M. A., Dixon, B. L., & Greene, J. P. (2012). Observed characteristics and teacher quality: Impacts of sample selection on a value added model. *Economics of Education Review,* 31(1), 19-32. doi.org/10.1016/j.econedurev.2011.07.014

5 Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, 12(3), 247-256. doi.org/10.1023/A:1008067210518

6 Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2014). The sensitivity of value-added estimates to specification adjustments: Evidence from school- and teacher-level models in Missouri. *Statistics and Public Policy,* 1(1), 19-27. doi.org/10.1080/2330443X.2013.856152

7 All statistics in this paragraph come from the 2017 *Digest of Education Statistics,* available online at https://nces.ed.gov/programs/digest/.

8 For more information, see Castellano & Ho, 2013. Op. cit.

9 Example and figure drawn from http://www.shankerinstitute.org/blog/when-growth-isnt-really-growth.

10 Welsh, R. O. (2017). School hopscotch: A comprehensive review of K–12 student mobility in the United States. *Review of Educational Research*, 87(3), 475-511. doi:10.3102/0034654316672068

11 CORE Districts. (2018). *Analysis of residual gain student-level growth models.* Sacramento, CA.

12 Data Quality Campaign. (2019). *Growth data: It matters, and it's complicated.* Washington, DC.

13 For a clear description of the similarities and differences between these two models, see Castellano and Ho, 2013. Op. cit.

14 Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review,* 47, 180-195. Op. cit.

15 See Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2014). Choosing the right growth measure. *Education Next,* 14(2), 67-72. Retrieved from https://www.educationnext.org/choosing-the-right-growth-measure/

16 Parsons, E., Koedel, C., & Tan, L. (2019). Accounting for student disadvantage in value-added models. *Journal of Educational and Behavioral Statistics,* 44(2), 144–179. doi:10.3102/1076998618803889

17 See Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill and Melinda Gates Foundation. See also Kane, T. J. & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. NBER Working Paper No. 14607.

18 Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review,* 104(9), 2593-2632. Retrieved from http://www.jstor.org/stable/43495327. Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2633-79. doi:10.3386/w19424.

19 Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review,* 47, 180-195. Op. cit.

20 Figure 2.1 from Hegedus, A. (2018). *Evaluating the relationships between poverty and school performance.* Portland, OR: NWEA.

21 Figure 2.2 from Hegedus, A. (2018). *Evaluating the relationships between poverty and school performance.* Portland, OR: NWEA.

22 Korn, S.A. (2018, March). *Would you recognize a quality school if you saw one?: Exploring parents' evaluation of schools using Mechanical Turk.* Paper session presented at the 44th annual meeting of the Association for Education Finance and Policy Conference, Kansas City, MO.

23 Education Trust. (2013). *The value of value-added data.* Washington, DC: Author. Tennessee SCORE. (2018). What teachers say about TVAAS. Nashville, TN: Author.

24 Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know.* Cambridge, MA: Harvard Education Press.

25 Castellano, K. E., & Ho, A.D. (2013). *A practitioner's guide to growth models.* Washington, DC: Council of Chief State School Officers.

26 Data Quality Campaign. (2019). *Growth data: It matters, and it's complicated.* Washington, DC: Author.

27 Castellano, K. E., & Ho, A.D. (2013). *A practitioner's guide to growth models.* Washington, DC: Council of Chief State School Officers. Op. cit.

28 Castellano, K. E., & Ho, A.D. (2013). A *practitioner's guide to growth models.* Washington, DC: Council of Chief State School Officers. Op. cit.

29 Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180-195. Op. Cit.

Policy Analysis for California Education (PACE)

Policy Analysis for California Education (PACE) is an independent, non-partisan research center led by faculty directors at Stanford University, the University of Southern California, the University of California Davis, the University of California Los Angeles, and the University of California Berkeley. PACE seeks to define and sustain a long-term strategy for comprehensive policy reform and continuous improvement in performance at all levels of California's education system, from early childhood to postsecondary education and training. PACE bridges the gap between research and policy, working with scholars from California's leading universities and with state and local policymakers to increase the impact of academic research on educational policy in California.

Founded in 1983, PACE

- Publishes policy briefs, research reports, and working papers that address key policy issues in California's education system.
- Convenes seminars and briefings that make current research accessible to policy audiences throughout California.
- Provides expert testimony on educational issues to legislative committees and other policy audiences.
- Works with local school districts and professional associations on projects aimed at supporting policy innovation, data use, and rigorous evaluation.

Recent Publications

Polikoff, M. S. (2019). *Gauging the revised California School Dashboard*. Palo Alto: Policy Analysis for California Education.

Phillips, M., Reber, S., & Rothstein, J. (2018). *Making California data more useful for educational improvement*. Palo Alto: Getting Down to Facts II, Policy Analysis for California Education and Stanford University.

Hough, H., Byun, E., & Mulfinger, L. (2018). *Using data for improvement: Learning from the CORE Data Collaborative*. Palo Alto: Getting Down to Facts II, Policy Analysis for California Education and Stanford University.

Koppich, J. E., White, E., Kim, S., Lauck, M., Bookman, N., & Venezia, A. (2019). *Developing a comprehensive data system to further continuous improvement in California.* Palo Alto: Policy Analysis for California Education.

Polikoff, M., Korn, S., & McFall, R. (2018). *In need of improvement? Assessing the California Dashboard after one year*. Palo Alto: Getting Down to Facts II, Policy Analysis for California Education and Stanford University.

PACE
Policy Analysis for California Education

Stanford Graduate School of Education
520 Galvez Mall, CERAS 401
Stanford, CA 94305-3001
Phone: (650) 724-2832
Fax: (650) 723-9931

edpolicyinca.org